Zhang, J & Katsaggelos, A.K. "Image Recovery Using the EM Algorithm"
Digital Signal Processing Handbook
Ed. Vijay K. Madisetti and Douglas B. Williams
Boca Raton: CRC Press LLC, 1999

29

Image Recovery Using the EM Algorithm

Jun Zhang
University of Wisconsin
Milwawkee

Aggelos K. Katsaggelos
Northwestern University

29.1 Introduction

Image recovery constitutes a significant portion of the inverse problems in image processing. Here, by image recovery we refer to two classes of problems, image restoration and image reconstruction. In image restoration, an estimate of the original image is obtained from a blurred and noise-corrupted image. In image reconstruction, an image is generated from measurements of various physical quantities, such as X-ray energy in CT and photon counts in single photon emission tomography (SPECT) and positron emission tomography (PET). Image restoration has been used to restore pictures in remote sensing, astronomy, medical imaging, art history studies, e.g., see [1], and more recently, it has been used to remove picture artifacts due to image compression, e.g., see [2] and [3]. While primarily used in biomedical imaging [4], image reconstruction has also found applications in materials studies [5].

Due to the inherent randomness in the scene and imaging process, images and noise are often best modeled as multidimensional random processes called random fields. Consequently, image recovery becomes the problem of *statistical inference*. This amounts to estimating certain unknown parameters of a probability density function (pdf) or calculating the expectations of certain random fields from the observed image or data. Recently, the *maximum-likelihood estimate* (MLE) has begun to play a central role in image recovery and led to a number of advances [6, 8]. The most significant advantage of the MLE over traditional techniques, such as the Wiener filtering, is perhaps that it can work more autonomously. For example, it can be used to restore an image with unknown blur and noise level by estimating them and the original image *simultaneously* [8, 9]. The traditional Wiener filter and other LMSE (least mean square error) techniques, on the other hand, would require the knowledge of the blur and noise level.

In the MLE, the likelihood function is the pdf evaluated at an observed data sample conditioned on the parameters of interest, e.g., blur filter coefficients and noise level, and the MLE seeks the parameters that maximize the likelihood function, i.e., best explain the observed data. Besides being intuitively appealing, the MLE also has several good asymptotic (large sample) properties [10] such as consistency (the estimate converges to the true parameters as the sample size increases). However, for many nontrivial image recovery problems, the direct evaluation of the MLE can be difficult, if not impossible. This difficulty is due to the fact that likelihood functions are usually highly nonlinear and often cannot be written in closed forms (e.g., they are often integrals of some other pdf's). While the former case would prevent analytic solutions, the latter case could make any numerical procedure impractical.

The EM algorithm, proposed by Dempster, Laird, and Rubin in 1977 [11], is a powerful iterative technique for overcoming these difficulties. Here, EM stands for *expectation-maximization*. The basic idea behind this approach is to introduce an auxiliary function (along with some auxiliary variables) such that it has similar behavior to the likelihood function but is much easier to maximize. By similar behavior, we mean that when the auxiliary function increases, the likelihood function also increases. Intuitively, this is somewhat similar to the use of auxiliary lines for the proofs in elementary geometry.

The EM algorithm was first used by Shepp and Verdi [7] in 1982 in emission tomography (medical imaging). It was first used by Katsaggelos and Lay [8] and Lagendijk et al. [9] for simultaneous image restoration and blur identification around 1989. The work of using the EM algorithm in image recovery has since flourished with impressive results. A recent search on the Compendex data base with key words "EM" and "image" turned up more than 60 journal and conference papers, published over the two and a half year period from January, 1993 to June, 1995.

Despite these successes, however, some fundamental problems in the application of the EM algorithm to image recovery remain. One is convergence. It has been noted that the estimates often do not converge, converge rather slowly, or converge to unsatisfactory solutions (e.g., spiky images) [12, 13]. Another problem is that, for some popular image models such as Markov random fields, the conditional expectation in the E-step of the EM algorithm can often be difficult to calculate [14]. Finally, the EM algorithm is rather general in that the choice of auxiliary variables and the auxiliary function is not unique. Is it possible that one choice is better than another with respect to convergence and expectation calculations [17]?

The purpose of this chapter is to demonstrate the application of the EM algorithm in some typical image recovery problems and survey the latest research work that addresses some of the fundamental problems described above. The chapter is organized as follows. In section 29.2, the EM algorithm is reviewed and demonstrated through a simple example. In section 29.3, recent work in convergence, expectation calculation, and the selection of auxiliary functions is discussed. In section 29.4, more complicated applications are demonstrated, followed by a summary in section 29.5. Most of the examples in this chapter are related to image restoration. This choice is motivated by two considerations — the mathematical formulations for image reconstruction are often similar to that of image restoration and a good account on image reconstruction is available in Snyder and Miller [6].

29.2 The EM Algorithm

Let the observed image or data in an image recovery problem be denoted by $\mathbf{y}$. Suppose that $\mathbf{y}$ can be modeled as a collection of random variables defined over a lattice $\mathbf{S}$ with $\mathbf{y} = \{y_i, i \in \mathbf{S}\}$. For example, $\mathbf{S}$ could be a square lattice of N^2 sites. Suppose that the pdf of $\mathbf{y}$ is $p_{\mathbf{y}}(\mathbf{y}|\theta)$, where θ is a set of parameters. In this chapter, $p(\cdot)$ is a general symbol for pdf and the subscript will be omitted whenever there is no confusion. For example, when $\mathbf{y}$ and $\mathbf{x}$ are two different random fields, their pdf's are represented as $p(\mathbf{y})$ and $p(\mathbf{x})$, respectively.

29.2.1 The Algorithm

Under statistical formulations, image recovery often amounts to seeking an estimate of θ, denoted by $\hat{\theta}$, from an observed $\mathbf{y}$. The MLE approach is to find $\hat{\theta}_{ML}$ such that

$$\hat{\theta}_{ML} = \arg\max_{\theta} p\left(\mathbf{y}|\theta\right) = \arg\max_{\theta} \log p\left(\mathbf{y}|\theta\right) \ , \tag{29.1}$$

where $p(\mathbf{y}|\theta)$, as a function of θ, is called the likelihood. As described previously, a direct solution of (29.1) can be difficult to obtain for many applications. The EM algorithm attempts to overcome this problem by introducing an auxiliary random field $\mathbf{x}$ with pdf $p(\mathbf{x}|\theta)$. Here, $\mathbf{x}$ is somewhat "more informative" [17] than $\mathbf{y}$ in that it is related to $\mathbf{y}$ by a *many-to-one* mapping

$$\mathbf{y} = \mathbf{H}(\mathbf{x}) \ . \tag{29.2}$$

That is, $\mathbf{y}$ can be regarded as a partial observation of $\mathbf{x}$, or *incomplete data*, with $\mathbf{x}$ being the *complete data*.

The EM algorithm attempts to obtain the incomplete data MLE of (29.1) through an iterative procedure. Starting with an initial estimate θ^0, each iteration k consists of two steps:

- *The E-step*: Compute the conditional expectation[1] $\langle \log p(\mathbf{x}|\theta)|\mathbf{y}, \theta^k\rangle$. This leads to a function of θ, denoted by $Q(\theta|\theta^k)$, which is the auxiliary function mentioned previously.
- *M-step*: Find θ^{k+1} from

$$\theta^{k+1} = \arg\max_{\theta} Q\left(\theta|\theta^k\right) \ . \tag{29.3}$$

It has been shown that the EM algorithm is monotonic [11], i.e., $\log p(\mathbf{y}|\theta^k) \geq \log p(\mathbf{y}|\theta^{k+1})$. It has also been shown that under mild regularity conditions, such as that the true θ must lie in the interior of a compact set and that the likelihood functions involved must have continuous derivatives, the estimate of θ from the EM algorithm converges, at least to a local maxima of $p(\mathbf{y}|\theta)$ [20, 21]. Finally, the EM algorithm extends easily to the case in which the MLE is used along with a penalty or a prior on θ. For example, suppose that $q(\theta)$ is a penalty to be minimized. Then, the M-step is modified to maximizing $Q(\theta|\theta^k) - q(\theta)$ with respect to θ.

29.2.2 Example: A Simple MRF

As an illustration of the EM algorithm, we consider a simple image restoration example. Let $\mathbf{S}$ be a two-dimensional square lattice. Suppose that the observed image $\mathbf{y}$ and the original image $\mathbf{u} = \{u_i, i \in \mathbf{S}\}$ are related through

$$\mathbf{y} = \mathbf{u} + \mathbf{w} \ , \tag{29.4}$$

where $\mathbf{w} = \{u_i, i \in \mathbf{S}\}$ is an i.i.d. additive zero-mean white Gaussian noise with variance σ^2. Suppose that $\mathbf{u}$ is modeled as a random field with an exponential or Gibbs pdf

$$p(\mathbf{u}) = Z^{-1} e^{-\beta E(\mathbf{u})} \tag{29.5}$$

---

[1] In this chapter, we use $\langle\cdot\rangle$ rather than $E[\cdot]$ to represent expectations since E is used to denote energy functions of the MRF.

where $E(\mathbf{u})$ is an *energy function* with

$$E(\mathbf{u}) = \frac{1}{2} \sum_i \sum_{j \in N_i} \phi\left(u_i, u_j\right) \tag{29.6}$$

and Z is a normalization factor

$$Z = \sum_{\mathbf{u}} e^{-\beta E(\mathbf{u})} \tag{29.7}$$

called the *partition function* whose evaluation generally involves all possible realizations of $\mathbf{u}$. In the energy function, N_i is a set of neighbors of i (e.g., the nearest four neighbors) and $\phi(\cdot, \cdot)$ is a nonlinear function called the *clique function*. The model for $\mathbf{u}$ is a simple but nontrivial case of the Markov random field (MRF) [22, 23] which, due to its versatility in modeling spatial interactions, has emerged as a powerful model for various image processing and computer vision applications [24].

A restoration that is optimal in the sense of minimum mean square error is

$$\hat{\mathbf{u}} = \langle \mathbf{u} | \mathbf{y} \rangle = \int \mathbf{u}\, p(\mathbf{u}|\mathbf{y})\, d\mathbf{u}\,. \tag{29.8}$$

If parameters β and σ^2 are known, the above expectation can be computed, at least approximately (see Conditional Expectation Calculations in section 29.3 for details). To estimate the parameters, now denoted by $\theta = (\beta, \sigma^2)$, one could use the MLE. Since $\mathbf{u}$ and $\mathbf{w}$ are independent,

$$p(\mathbf{y}|\theta) = \int p_{\mathbf{u}}(\mathbf{v}|\theta)\, p_{\mathbf{w}}(\mathbf{y} - \mathbf{v}|\theta)\, d\mathbf{v} = \left(p_{\mathbf{u}} * p_{\mathbf{w}}\right)(\mathbf{y}|\theta)\,, \tag{29.9}$$

where $*$ denotes convolution, and we have used some subscripts to avoid ambiguity. Notice that the integration involved in the convolution generally does not have a closed-form expression. Furthermore, for most types of clique functions, Z is a function of β and its evaluation is exponentially complex. Hence, direct MLE does not seem possible.

To try with the EM algorithm, we first need to select the complete data. A natural choice here, for example, is to let

$$\mathbf{x} = (\mathbf{u}, \mathbf{w}) \tag{29.10}$$

$$\mathbf{y} = \mathbf{H}(\mathbf{x}) = \mathbf{H}(\mathbf{u}, \mathbf{w}) = \mathbf{u} + \mathbf{w}\,. \tag{29.11}$$

Clearly, many different $\mathbf{x}$ can lead to the same $\mathbf{y}$. Since $\mathbf{u}$ and $\mathbf{w}$ are independent, $p(\mathbf{x}|\theta)$ can be found easily as

$$p(\mathbf{x}|\theta) = p(\mathbf{u})\, p(\mathbf{w})\,. \tag{29.12}$$

However, as the reader can verify, one encounters difficulty in the derivation of $p(\mathbf{x}|\mathbf{y}, \theta^k)$ which is needed for the conditional expectation of the E-step. Another choice is to let

$$\mathbf{x} = (\mathbf{u}, \mathbf{y}) \tag{29.13}$$

$$\mathbf{y} = H(\mathbf{u}, \mathbf{y}) = \mathbf{y} \tag{29.14}$$

The log likelihood of the complete data is

$$\begin{aligned} \log p(\mathbf{x}|\theta) &= \log p(\mathbf{y}, \mathbf{u}|\theta) \\ &= \log p(\mathbf{y}|\mathbf{u}, \theta)\, p(\mathbf{u}|\theta) \\ &= c - \sum_i \frac{(y_i - u_i)^2}{2\sigma^2} - \log Z(\beta) - \frac{\beta}{2} \sum_i \sum_{j \in N_i} \phi\left(u_i, u_j\right)\,, \end{aligned} \tag{29.15}$$

where c is a constant. From this we see that in the E-step, we only need to calculate three types of terms, $\langle u_i \rangle$, $\langle u_i^2 \rangle$, and $\langle \phi(u_i, u_j) \rangle$. Here, the expectations are all conditioned on $\mathbf{y}$ and θ^k. To compute these expectations, one needs the conditional pdf $p(\mathbf{u}|\mathbf{y}, \theta^k)$ which is, from Bayes' formula,

$$
\begin{aligned}
p\left(\mathbf{u}|\mathbf{y}, \theta^k\right) &= \frac{p\left(\mathbf{y}|\mathbf{u}, \theta^k\right) p\left(\mathbf{u}|\theta^k\right)}{p\left(\mathbf{y}|\theta^k\right)} \\
&= \left[2\pi\sigma^2\right]^{-||\mathbf{S}||/2} e^{-\sum_i (y_i - u_i)^2/2\left(\sigma^2\right)^k} Z^{-1} e^{-\beta^k E(\mathbf{u})} \left[p\left(\mathbf{y}|\theta^k\right)\right]^{-1} . \qquad (29.16)
\end{aligned}
$$

Here, the superscript k denotes the kth iteration rather than the kth power. Combining all the constants and terms in the exponentials, the above equation becomes that of a Gibbs distribution

$$
p\left(\mathbf{u}|\mathbf{y}, \theta^k\right) = Z_1^{-1}\left(\theta^k\right) e^{-E_1\left(\mathbf{u}|\mathbf{y}, \theta^k\right)} \qquad (29.17)
$$

where the energy function is

$$
E_1\left(\mathbf{u}|\mathbf{y}, \theta^k\right) = \sum_i \left[\frac{(y_i - u_i)^2}{2\left(\sigma^2\right)^k} + \frac{\beta^k}{2} \sum_{j \in N_i} \phi\left(u_i, u_j\right)\right] . \qquad (29.18)
$$

Even with this, the computation of the conditional expectation in the E-step can still be a difficult problem due to the coupling of the u_i and u_j in E_1. This is one of the fundamental problems of the EM algorithm that will be addressed in section 29.3. For the moment, we assume that the E-step can be performed successfully with

$$
\begin{aligned}
Q\left(\theta|\theta^k\right) &= \langle \log p(\mathbf{x}|\theta)|\mathbf{y}, \theta^k \rangle \\
&= c - \sum_i \frac{\langle (y_i - x_i)^2 \rangle^k}{2\sigma^2} - \log Z(\beta) - \frac{\beta}{2} \sum_i \sum_{j \in N_i} \langle \phi\left(u_i, u_j\right) \rangle^k , \qquad (29.19)
\end{aligned}
$$

where $\langle \cdot \rangle^k$ is an abbreviation for $\langle \cdot |\mathbf{y}, \theta^k \rangle$. In the M-step, the update for θ can be found easily by setting

$$
\frac{\partial}{\partial \sigma^2} Q\left(\theta|\theta^k\right) = 0 , \qquad \frac{\partial}{\partial \beta} Q\left(\theta|\theta^k\right) = 0 . \qquad (29.20)
$$

From the first of these,

$$
\left(\sigma^2\right)^{k+1} = ||\mathbf{S}||^{-1} \sum_i \langle (y_i - u_i)^2 \rangle^k \qquad (29.21)
$$

The solution of the second equation, on the other hand, is generally difficult due to the well-known difficulties of evaluating the partition function $Z(\beta)$ (see also Eq. (29.7)) which needs to be dealt with via specialized approximations [22, 25]. However, as demonstrated by Bouman and Sauer [26], some simple yet important cases exist in which the solution is straightforward. For example, when $\phi(u_i, u_j) = (u_i - u_j)^2$, $Z(\beta)$ can be written as

$$
\begin{aligned}
Z(\beta) &= \int e^{-\frac{\beta}{2} \sum_i \sum_{j \in N_i} \left(u_i - u_j\right)^2} d\mathbf{u} \\
&= \beta^{-||\mathbf{S}||/2} \int e^{-\frac{1}{2} \sum_i \sum_{j \in N_i} \left(v_i - v_j\right)^2} d\mathbf{v} = \beta^{-||\mathbf{S}||/2} Z(1) . \qquad (29.22)
\end{aligned}
$$

Here, we have used a change of variable, $v_i = \sqrt{\beta} u_i$. Now, the update of β can be found easily as

$$
\beta^{k+1} = ||\mathbf{S}||^{-1} \sum_i \sum_{j \in N_i} \langle \left(u_i - u_j\right)^2 \rangle^k . \qquad (29.23)
$$

This simple technique applies to a wider class of clique functions characterized by $\phi(u_i, u_j) = |u_i - u_j|^r$ with any $r > 0$ [26].

29.3 Some Fundamental Problems

As is in many other areas of signal processing, the power and versatility of the EM algorithm has been demonstrated in a large number of diverse image recovery applications. Previous work, however, has also revealed some of its weaknesses. For example, the conditional expectation of the E-step can be difficult to calculate analytically and too time-consuming to compute numerically, as is in the MRF example in the previous section. To a lesser extent, similar remarks can be made to the M-step. Since the EM algorithm is iterative, convergence can often be a problem. For example, it can be very slow. In some applications, e.g., emission tomography, it could converge to the wrong result — the reconstructed image gets spikier as the number of iterations increases [12, 13]. While some of these problems, such as slow convergence, are common to many numerical algorithms, most of their causes are inherent to the EM algorithm [17, 19].

In previous work, the EM algorithm has mostly been applied in a "natural fashion" (e.g., in terms of selecting incomplete and complete data sets) and the problems mentioned above were dealt with on an ad hoc basis with mixed results. Recently, however, there has been interest in seeking more fundamental solutions [14, 19]. In this section, we briefly describe the solutions to two major problems related to the EM algorithm, namely, the conditional expectation computation in the E-step when the data is modeled as MRF's and fundamental ways of improving convergence.

29.3.1 Conditional Expectation Calculations

When the complete data is an MRF, the conditional expectation of the E-step of the EM algorithm can be difficult to perform. For instance, consider the simple MRF in section 29.2, where it amounts to calculating $\langle u_i \rangle$, $\langle u_i^2 \rangle$, and $\langle \phi(u_i, u_j) \rangle$ and the expectations are taken with respect to $p(\mathbf{u}|\mathbf{y}, \theta^k)$ of Eq. (29.17). For example, we have

$$\langle u_i \rangle = Z_1^{-1} \int u_i e^{-E_1(\mathbf{u})} \, d\mathbf{u} \tag{29.24}$$

Here, for the sake of simplicity, we have omitted the superscript k and the parameters, and this is done in the rest of this section whenever there is no confusion. Since the variables u_i and u_j are coupled in the energy function for all i and j that are neighbors, the pdf and Z_1 cannot be factored into simpler terms, and the integration is exponentially complex, i.e., it involves all possible realizations of $\mathbf{u}$. Hence, some approximation scheme has to be used. One of these is the Monte Carlo simulation. For example, Gibbs samplers [23] and Metropolis techniques [27] have been used to generate samples according to $p(\mathbf{u}|\mathbf{y}, \theta^k)$ [26, 28]. A disadvantage of these is that, generally, hundreds of samples of $\mathbf{u}$ are needed and if the image size is large, this can be computation intensive. Another technique is based on the mean field theory (MFT) of statistical mechanics [25]. This has the advantage of being computationally inexpensive while providing satisfactory results in many practical applications. In this section, we will outline the essentials of this technique.

Let $\mathbf{u}$ be an MRF with pdf

$$p(\mathbf{u}) = Z^{-1} e^{-\beta E(\mathbf{u})} \,. \tag{29.25}$$

For the sake of simplicity, we assume that the energy function is of the form

$$E(\mathbf{u}) = \sum_i \left[h_i(u_i) + \frac{1}{2} \sum_{j \in N_i} \phi\left(u_i, u_j\right) \right] \tag{29.26}$$

where $h_i(\cdot)$ and $\phi(\cdot,\cdot)$ are some suitable, and possibly nonlinear, functions. The mean field theory attempts to derive a pdf $p_{MF}(\mathbf{u})$ that is an approximation to $p(\mathbf{u})$ and can be factored like an independent pdf.

The MFT used previously can be divided into two classes, the local mean field energy (LMFE) and the ones based on the Gibbs-Bogoliubov-Feynman (GBF) inequality. The LMFE scheme is based on the idea that when calculating the mean of the MRF at a given site, the influence of the random variables at other sites can be approximated by the influence of their means. Hence, if we want to calculate the mean of u_i, a *local energy function* can be constructed by collecting all the terms in (29.26) that are related to u_i and replacing the u_j's by their mean. Hence, for this energy function we have

$$E_i^{MF}(u_i) = h_i(u_i) + \sum_{i\in N_i} \phi\left(u_i, \langle u_j\rangle\right) \tag{29.27}$$

$$p_i^{MF}(u_i) = Z_i^{-1} e^{-\beta E_i^{MF}(u_i)} \tag{29.28}$$

$$p_{MF}(\mathbf{u}) = \prod_i p_i^{MF}(u_i) \tag{29.29}$$

Using this mean field pdf, the expectation of u_i and its functions can be found easily.

Again we use the MRF example from section 29.2.2 as an illustration. Its energy function is (29.18) and for the sake of simplicity, we assume that $\phi(u_i, u_j) = |u_i - u_j|^2$. By the LMFE scheme,

$$E_i^{MF} = \frac{(y_i - u_i)^2}{2\sigma^2} + \sum_{j\in N_i} \beta\left(u_i - \langle u_j\rangle\right)^2 \tag{29.30}$$

which is the energy of a Gaussian. Hence, the mean can be found easily by completing the square in (29.30) with

$$\langle u_i\rangle = \frac{y_i/\sigma^2 + 2\beta\sum_{j\in N_i}\langle u_j\rangle}{1/\sigma^2 + 2\beta||N_i||}\,. \tag{29.31}$$

When $\phi(\cdot,\cdot)$ is some general nonlinear function, numerical integration might be needed. However, compared to (29.24) such integrals are all with respect to one or two variables and are easy to compute.

Compared to the physically motivated scheme above, the GBF is an optimization approach. Suppose that $p_0(\mathbf{u})$ is a pdf which we want to use to approximate another pdf, $p(\mathbf{u})$. According to information theory, e.g., see [29], the *directed-divergence* between p_0 and p is defined as

$$D(p_0||p) = \langle \log p_0(\mathbf{u}) - \log p(\mathbf{u})\rangle_0, \tag{29.32}$$

where the subscript 0 indicates that the expectation is taken with respect to p_0, and it satisfies

$$D(p_0||p) \geq 0 \tag{29.33}$$

with equality holds if and only if $p_0 = p$. When the pdf's are Gibbs distributions, with energy functions E_0 and E and partition functions Z_0 and Z, respectively, the inequality becomes

$$\log Z \geq \log Z_0 - \beta\langle E - E_0\rangle_0 = \log Z_0 - \beta\langle \Delta E\rangle_0\,, \tag{29.34}$$

which is known as the GBF inequality.

Let p_0 be a parametric Gibbs pdf with a set of parameters ω to be determined. Then, one can obtain an optimal p_0 by maximizing the right-hand side of (29.34). As an illustration, consider again the MRF example in section 29.2 with the energy function (29.18) and a quadratic clique function, as we did for the LMFE scheme. To use the GBF, let the energy function of p_0 be defined as

$$E_0(\mathbf{u}) = \sum_i \frac{(u_i - m_i)^2}{2v_i^2} \tag{29.35}$$

where $\{m_i, \nu_i^2, i \in \mathbf{S}\} = \omega$ is the set of parameters to be determined in the maximization of the GBF. Since this is the energy for an independent Gaussian, Z_0 is just

$$Z_0 = \prod_i \sqrt{2\pi \nu_i^2}. \tag{29.36}$$

The parameters of p_0 can be obtained by finding an expression for the right-hand side of the GBF inequality, letting its partial derivatives (with respect to the parameters m_i and ν_i^2) be zero, and solving for the parameters. Through a somewhat lengthy but straightforward derivation, one can find that [30]

$$m_i = \frac{y_i/\sigma^2 + 2\beta \sum_{j \in N_i} \langle u_j \rangle}{1/\sigma^2 + 2\beta ||N_i||} \ . \tag{29.37}$$

Since $m_i = \langle u_i \rangle$, the GBF produces the same result as the LMEF. This, however, is an exception rather than the rule [30] and it is due to the quadratic structures of both energy functions.

We end this section with several remarks. First, compared to the LMFE, the GBF scheme is an optimization scheme, hence more desirable. However, if the energy function of the original pdf is highly nonlinear, the GBF could require the solution of a difficult nonlinear equation in many variables (see e.g., [30]). The LMFE, though not optimal, can always be implemented relatively easily. Secondly, while the MFT techniques are significantly more computation-efficient than the Monte Carlo techniques and provide good results in many applications, no proof exists as yet that the conditional mean computed by the MFT will converge to the true conditional mean. Finally, the performance of the mean field approximations may be improved by using "high-order" models. For example, one simple scheme is to consider LMFE's with a pair of neighboring variables [25, 31]. For the energy function in (29.26), for example, the "second-order" LMFE is

$$E_{i,j}^{MF}(u_i, u_j) = h_i(u_i) + h_i(u_j) + \beta \sum_{i' \in N_i} \phi(u_i, \langle u_{i'} \rangle) + \beta \sum_{j' \in N_j} \phi(u_j, \langle u_{j'} \rangle) \tag{29.38}$$

and

$$p_{MF}(u_i, u_j) = Z_{MF}^{-1} e^{-\beta E_{i,j}^{MF}(u_i, u_j)}, \tag{29.39}$$

$$p_{MF}(u_i) = \int p_{MF}\left(u_i, u_j\right) du_j \ . \tag{29.40}$$

Notice that (29.40) is not the same as (29.28) in that the fluctuation of u_j is taken into consideration.

29.3.2 Convergence Problem

Research on the EM algorithm-based image recovery has so far suggested two causes for the convergence problems mentioned previously. The first is whether the random field models used adequately capture the characteristics and constraints of the underlying physical phenomenon. For example, in emission tomography the original EM procedure of Shepp and Verdi tends to produce spikier and spikier images as the number of iteration increases [13]. It was found later that this is due to the assumption that the densities of the radioactive material at different spatial locations are independent. Consequently, various smoothness constraints (density dependence between neighboring locations) have been introduced as penalty functions or priors and the problem has been greatly reduced. Another example is in blind image restoration. It has been found that in order for the EM algorithm to produce reasonable estimate of the blur, various constraints need to be imposed. For instance, symmetry conditions and good initial guesses (e.g., a lowpass filter) are used in [8] and [9]. Since the blur tends to have a smooth impulse response, orthonormal expansion (e.g., the DCT) has also been used to reduce ("compress") the number of parameters in its representation [15].

The second factor that can be quite influential to the convergence of the EM algorithm, noticed earlier by Feder and Weinstein [16], is how the complete data is selected. In their work [18], Fessler and Hero found that for some EM procedures, it is possible to significantly increase the convergence rate by properly defining the complete data. Their idea is based on the observation that the EM algorithm, which is essentially a MLE procedure, often converges faster if the parameters are estimated sequentially in small groups rather than simultaneously. Suppose, for example, that 100 parameters are to be estimated. It is much better to estimate, in each EM cycle, the first 10 while holding the next 90 constant; then estimate the next 10 holding the remaining 80 and the newly updated 10 parameters constant; and so on. This type of algorithm is called the SAGE (Space Alternating Generalized EM) algorithm.

We illustrate this idea through a simple example used by Fessler and Hero [18]. Consider a simple image recovery problem, modeled as

$$\mathbf{y} = \mathbf{A}_1\theta_1 + \mathbf{A}_2\theta_2 + \mathbf{n} \,. \tag{29.41}$$

Column vectors θ_1 and θ_2 represent two original images or two data sources, $\mathbf{A}_1$ and $\mathbf{A}_2$ are two blur functions represented as matrices, and $\mathbf{n}$ is an additive white Gaussian noise source. In this model, the observed image $\mathbf{y}$ is the noise-corrupted combination of two blurred images (or data sources). A natural choice for the complete data is to view $\mathbf{n}$ as the combination of two smaller noise sources, each associated with one original image, i.e.,

$$\mathbf{x} = [\mathbf{A}_1\theta_1 + \mathbf{n_1}, \mathbf{A}_2\theta_2 + \mathbf{n}_2]' \,. \tag{29.42}$$

where $\mathbf{n}_1$ and $\mathbf{n}_2$ are i.i.d additive white Gaussian noise vectors with covariance matrix $\frac{\sigma^2}{2}\mathbf{I}$ and $'$ denotes transpose. The incomplete data $\mathbf{y}$ can be obtained from $\mathbf{x}$ by

$$\mathbf{y} = [\mathbf{I}, \mathbf{I}]\mathbf{x} \,. \tag{29.43}$$

Notice that this is a Gaussian problem in that both $\mathbf{x}$ and $\mathbf{y}$ are Gaussian and they are jointly Gaussian as well. From the properties of jointly Gaussian random variables [32], the EM cycle can be found relatively straightforwardly as

$$\theta_1^{k+1} = \theta_1^k + (\mathbf{A}_1'\mathbf{A}_1)^{-1}\mathbf{A}_1'\hat{\epsilon}/2\sigma^2 \tag{29.44}$$

$$\theta_2^{k+1} = \theta_2^k + (\mathbf{A}_2'\mathbf{A}_2)^{-1}\mathbf{A}_2'\hat{\epsilon}/2\sigma^2 \tag{29.45}$$

where

$$\hat{\epsilon} = (\mathbf{y} - \mathbf{A}_1\theta_1^k - \mathbf{A}_2\theta_2^k)/\sigma^2 \tag{29.46}$$

The SAGE algorithm for this simple problem is obtained by defining two smaller "complete data sets",

$$\mathbf{x}_1 = \mathbf{A}_1\theta_1 + \mathbf{n} \,, \qquad \mathbf{x}_2 = \mathbf{A}_2\theta_2 + \mathbf{n} \,. \tag{29.47}$$

Notice that now the noise $\mathbf{n}$ is associated "totally" with each smaller complete data set. The incomplete data $\mathbf{y}$ can be obtained from both $\mathbf{x}_1$ and $\mathbf{x}_2$, e.g.,

$$\mathbf{y} = \mathbf{x}_1 + \mathbf{A}_2\theta_2 \tag{29.48}$$

The SAGE algorithm amounts to two sequential and "smaller" EM algorithms. Specifically, corresponding to each classical EM cycle (29.44)-(29.46), the first SAGE cycle is a classical EM cycle with $\mathbf{x}_1$ as the complete data and θ_1 as the parameter set to be updated. The second SAGE cycle is a classical EM cycle with $\mathbf{x}_2$ as the complete data and θ_2 as the parameter set to be updated. The new update of θ_1 is also used. The specific algorithm is

$$\theta_1^{k+1} = \theta_1^k + \left(\mathbf{A}_1'\mathbf{A}_1\right)^{-1}\mathbf{A}_1'\hat{\epsilon}_1/2\sigma^2 \tag{29.49}$$

$$\theta_2^{k+1} = \theta_2^k + \left(\mathbf{A}_2'\mathbf{A}_2\right)^{-1}\mathbf{A}_2'\hat{\epsilon}_2/2\sigma^2 \tag{29.50}$$

where

$$\hat{\epsilon}_1 = \left(\mathbf{y} - \mathbf{A}_1\theta_1^k - \mathbf{A}_2\theta_2^k\right)/\sigma^2 \tag{29.51}$$

$$\hat{\epsilon}_2 = \left(\mathbf{y} - \mathbf{A}_1\theta_1^{k+1} - \mathbf{A}_2\theta_2^k\right)/\sigma^2 \tag{29.52}$$

We end this subsection with several remarks. First, for a wide class of random field models including the simple one above, Fessler and Hero have shown that the SAGE converges significantly faster than the classical EM [17]. In some applications, e.g., tomography, an acceleration of 5 to 10 times may be achieved. Secondly, just as for the EM algorithm, various constraints on the parameters are often needed and can be imposed easily as penalty functions in the SAGE algorithm. Finally, notice that in (29.41), the original images are treated as parameters (with constraints) rather than as random variables with their own pdfs. It would be of interest to investigate a Bayesian counterpart of the SAGE algorithm.

29.4 Applications

In this section, we describe the application of the EM algorithm to the simultaneous identification of the blur and image model and the restoration of single and multichannel images.

29.4.1 Single Channel Blur Identification and Image Restoration

Most of the work on restoration in the literature was done under the assumption that the blurring process (usually modeled as a linear space-invariant system (LSI) and specified by its point spread function (PSF)) is exactly known (for recent reviews of the restoration work in the literature see [8, 33]). However, this may not be the case in practice since usually we do not have enough knowledge about the mechanism of the degradation process. Therefore, the estimation of the parameters that characterize the degradation operator needs to be based on the available noisy and blurred data.

Problem formulation

The observed image $y(i, j)$ is modeled as the output of a 2D LSI system with PSF $\{d(p, q)\}$. In the following we will use (i, j) to denote a location on the lattice $\mathbf{S}$, instead of a single subscript. The output of the LSI system is corrupted by additive zero-mean Gaussian noise $v(i, j)$ with covariance matrix $\Lambda_{\mathbf{V}}$, which is uncorrelated with the original image $u(i, j)$. That is, the observed image $y(i, j)$ is expressed as

$$y(i, j) = \sum_{(p,q)\in \mathbf{S_D}} d(p, q)u(i - p, j - q) + v(i, j)\ , \tag{29.53}$$

where $\mathbf{S_D}$ is the finite support region of the distortion filter. We assume that the arrays $y(i, j)$, $u(i, j)$, and $v(i, j)$ are of size $N \times N$. By stacking them into $N^2 \times 1$ vectors, Eq. (29.53) can be rewritten in matrix/vector form as [35]

$$\mathbf{y} = \mathbf{D}\mathbf{u} + \mathbf{v}\ , \tag{29.54}$$

where $\mathbf{D}$ is an $N^2 \times N^2$ matrix.

The vector $\mathbf{u}$ is modeled as a zero-mean Gaussian random field. Its pdf is equal to

$$p(\mathbf{u}) = |2\pi\Lambda_{\mathbf{U}}|^{-\mathbf{1/2}} \exp\left\{\frac{-\mathbf{1}}{\mathbf{2}}\mathbf{u}^{\mathbf{H}}\Lambda_{\mathbf{U}}{}^{-\mathbf{1}}\mathbf{u}\right\}\ , \tag{29.55}$$

where $\Lambda_{\mathbf{U}}$ is the covariance matrix of $\mathbf{u}$, ^{H} denotes the Hermitian (i.e. conjugate transpose) of a matrix and a vector, and $|\cdot|$ denotes the determinant of a matrix. A special case of this representation is when $u(i, j)$ is described by an autoregressive (AR) model. Then $\Lambda_{\mathbf{U}}$ can be parameterized in terms of the AR coefficients and the covariance of the driving noise [38, 57].

Equation (29.53) can be written in the continuous frequency domain according to the convolution theorem. Since the discrete Fourier transform (DFT) will be used in implementing convolution, we assume that Eq. (29.53) represents circular convolution (2D sequences can be padded with zeros in such a way that the result of the linear convolution equals that of the circular convolution, or the observed image can be preprocessed around its boundaries so that Eq. (29.53) is consistent with the circular convolution of $\{d(p, q)\}$ with $\{u(p, q)\}$ [36]). Matrix $\mathbf{D}$ then becomes block circulant [35].

Maximum Likelihood (ML) Parameter Identification

The assumed image and blur models are specified in terms of the deterministic parameters $\theta = \{\Lambda_{\mathbf{U}}, \Lambda_{\mathbf{V}}, \mathbf{D}\}$. Since $\mathbf{u}$ and $\mathbf{v}$ are uncorrelated, the observed image $\mathbf{y}$ is also Gaussian with pdf equal to

$$\begin{aligned} p(\mathbf{y}/\theta) &= |2\pi \left(\mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^{\mathbf{H}} + \Lambda_{\mathbf{V}}\right)|^{-1/2} \\ & \exp\left\{\frac{-1}{2}\mathbf{y}^T \left(\mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^{\mathbf{H}} + \Lambda_{\mathbf{V}}\right)^{-1} \mathbf{y}\right\}, \end{aligned} \tag{29.56}$$

where the inverse of the matrix $(\mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^{\mathbf{H}} + \Lambda_{\mathbf{V}})$ is assumed to be defined since covariance matrices are symmetric positive definite.

Taking the logarithm of Eq. (29.56) and disregarding constant additive and multiplicative terms, the maximization of the log-likelihood function becomes the minimization of the function $L(\theta)$, given by

$$L(\theta) = \log|\mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^{\mathbf{H}} + \Lambda_{\mathbf{V}}| + \left\{\mathbf{y}^{\mathbf{T}} \left(\mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^{\mathbf{H}} + \Lambda_{\mathbf{V}}\right)^{-\mathbf{1}} \mathbf{y}\right\}. \tag{29.57}$$

By studying the function $L(\theta)$ it is clear that if no structure is imposed on the matrices $\mathbf{D}$, $\Lambda_{\mathbf{U}}$, and $\Lambda_{\mathbf{V}}$, the number of unknowns involved is very large. With so many unknowns and only one observation (i.e., $\mathbf{y}$), the ML identification problem becomes unmanageable. Furthermore, the estimate of $\{d(p, q)\}$ is not unique, because the ML approach to image and blur identification uses only second order statistics of the blurred image, since all pdfs are assumed to be Gaussian. More specifically, the second order statistics of the blurred image do not contain information about the phase of the blur, which, therefore, is in general undetermined. In order to restrict the set of solutions and hopefully obtain a unique solution, additional information about the unknown parameters needs to be incorporated into the solution process.

The structure we are imposing on $\Lambda_{\mathbf{U}}$ and $\Lambda_{\mathbf{V}}$ results from the commonly used assumptions in the field of image restoration [35]. First we assume that the additive noise $\mathbf{v}$ is white, with variance $\sigma_{\mathbf{V}}^2$, that is,

$$\Lambda_{\mathbf{V}} = \sigma_{\mathbf{V}}^{\mathbf{2}}\mathbf{I}\,. \tag{29.58}$$

Further we assume that the random process $\mathbf{u}$ is stationary which results in $\Lambda_{\mathbf{U}}$ being a block Toeplitz matrix [35]. A block Toeplitz matrix is asymptotically equivalent to a block circulant matrix as the dimension of the matrix becomes large [37]. For average size images, the dimensions of $\Lambda_{\mathbf{U}}$ are large indeed; therefore, the block circulant approximation is a valid one. Associated with $\Lambda_{\mathbf{U}}$ are the 2D sequences $\{l_{\mathbf{U}}(p, q)\}$. The matrix $\mathbf{D}$ in Eq. (29.54) was also assumed to be block circulant. Block circulant matrices can be diagonalized with a transformation matrix constructed from discrete Fourier kernels [35]. The diagonal matrices corresponding to $\Lambda_{\mathbf{U}}$ and $\mathbf{D}$ are denoted respectively by $\mathbf{Q}_{\mathbf{U}}$ and $\mathbf{Q}_{\mathbf{D}}$. They have as elements the raster scanned 2D DFT values of the 2D sequences $\{l_{\mathbf{U}}(p, q)\}$ and $\{d(p, q)\}$, denoted respectively by $S_{\mathbf{U}}(m, n)$ and $\Delta(m, n)$.

Due to the above assumptions Eq. (29.57) can be written in the frequency domain as

$$L(\theta) = \sum_{m=0}^{N-1} \sum_{n=0}^{N-1} \left\{ \log \left(|\Delta(m,n)|^2 S_{\mathbf{U}}(m,n) + \sigma_{\mathbf{V}}^2 \right) + \frac{|Y(m,n)|^2}{|\Delta(m,n)|^2 S_{\mathbf{U}}(m,n) + \sigma_{\mathbf{V}}^2} \right\} , \tag{29.59}$$

where $Y(m,n)$ is the 2D DFT of $y(i,j)$. Equation (29.59) more clearly demonstrates the already mentioned nonuniqueness of the ML blur solution, since only the magnitude of $\Delta(m,n)$ appears in $L(\theta)$. If the blur is zero-phase, as is the case with $\mathbf{D}$ modeling atmospheric turbulence with long exposure times and mild defocussing ($\{d(p,q)\}$ is 2D Gaussian in this case), then a unique solution may be obtained. Nonuniqueness of the estimation of $\{d(p,q)\}$ can in general be avoided by enforcing the solution to satisfy a set of constraints. Most PSFs of practical interest can be assumed to be symmetric, i.e., $d(p,q) = d(-p,-q)$. In this case the phase of the DFT of $\{d(p,q)\}$ is zero or $\pm\pi$. Unfortunately, uniqueness of the ML solution is not always established by the symmetry assumption, due primarily to the phase ambiguity. Therefore, additional constraints may alleviate this ambiguity. Such additional constraints are the following: (1) The PSF coefficients are nonnegative, (2) the support $\mathbf{S_D}$ is finite, and (3) the blurring mechanism preserves energy [35], which results in

$$\sum_{(i,j)\in \mathbf{S_D}} d(i,j) = 1 . \tag{29.60}$$

The EM Iterations for the ML Estimation of θ

The next step to be taken in implementing the EM algorithm is the determination of the mapping $\mathbf{H}$ in Eq. (29.2). Clearly Eq. (29.54) can be rewritten as

$$\mathbf{y} = \begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{y} \end{bmatrix} = \begin{bmatrix} \mathbf{D} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{Du} \\ \mathbf{v} \end{bmatrix} , \tag{29.61}$$

where $\mathbf{0}$ and $\mathbf{I}$ represent the $N^2 \times N^2$ zero and identity matrices, respectively. Therefore, according to Eq. (29.61), there are three candidates for representing the complete data $\mathbf{x}$, namely, $\{\mathbf{u},\mathbf{y}\}$, $\{\mathbf{u},\mathbf{v}\}$, and $\{\mathbf{Du},\mathbf{v}\}$. All three cases are analyzed in the following. However, as it will be shown, only the choice of $\{\mathbf{u},\mathbf{y}\}$ as the complete data fully justifies the term "complete data", since it results in the simultaneous identification of all unknown parameters and the restoration of the image.

For the case when $\mathbf{H}$ in Eq. (29.2) is linear, as are the cases represented by Eq. (29.61), and the data $\mathbf{y}$ is modeled as a zero-mean Gaussian process, as is the case under consideration expressed by Eq. (29.56), the following general result holds for all three choices of the complete data [38, 39, 57].

The E-step of the algorithm results in the computation of $Q(\theta/\theta^k)$ =constant$-F(\theta/\theta^k)$ where

$$\begin{aligned} F(\theta/\theta^k) &= \log|\Lambda_{\mathbf{X}}| + tr\left(\Lambda_{\mathbf{X}}^{-1}\mathbf{C}_{\mathbf{X|y}}^k\right) \\ &= \log|\Lambda_{\mathbf{X}}| + tr\left(\Lambda_{\mathbf{X}}^{-1}\Lambda_{\mathbf{X|y}}^k\right) + \mu_{\mathbf{X|y}}^{(k)H}\Lambda_{\mathbf{X}}^{-1}\mu_{\mathbf{X|y}}^k , \end{aligned} \tag{29.62}$$

where $\Lambda_{\mathbf{X}}$ is the covariance of the complete data $\mathbf{x}$ which is also a zero-mean Gaussian process,

$$\begin{aligned} \mathbf{C}_{\mathbf{X|y}}^k &= \langle \mathbf{xx}^H|\mathbf{y};\quad \theta^k\rangle = \Lambda_{\mathbf{X|y}}^k + \mu_{X|y}^k \mu_{\mathbf{X|y}}^{(k)H} , \\ \mu_{\mathbf{X|y}}^k &= \langle \mathbf{x}|\mathbf{y};\quad \theta^k\rangle = \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{Y}}^{-1}\mathbf{y} = \Lambda_{\mathbf{X}}\mathbf{H}^H\left(\mathbf{H}\Lambda_{\mathbf{XH}}^H\right)^{-1}\mathbf{y} , \end{aligned} \tag{29.63}$$

and

$$\Lambda_{\mathbf{X}|\mathbf{y}} = \langle(\mathbf{x}-\mu_{\mathbf{X}|\mathbf{y}})(\mathbf{x}-\mu_{\mathbf{X}|\mathbf{y}})^H|\mathbf{y};\theta^k\rangle = \Lambda_{\mathbf{X}} - \Lambda_{\mathbf{XY}}\Lambda_{\mathbf{Y}}^{-1}\Lambda_{\mathbf{YX}} = \Lambda_{\mathbf{X}} - \Lambda_{\mathbf{X}}\mathbf{H}^H\left(\mathbf{H}\Lambda_{\mathbf{X}}\mathbf{H}^H\right)^{-1}\mathbf{H}\Lambda_{\mathbf{X}}\,. \tag{29.64}$$

The M-step of the algorithm is described by the following equation

$$\theta^{(k+1)} = arg\left\{\min_{\{\theta\}} F(\theta/\theta^k)\right\}\,. \tag{29.65}$$

In our formulation of the identification/restoration problem the original image is not one of the unknown parameters in the set θ. However, as it will be shown in the next section, the restored image will be obtained in the E-step of the iterative algorithm.

{ u,y} as the complete data (CD_uy algorithm)

Choosing the original and observed images as the complete data, we obtain $\mathbf{H} = [\mathbf{0}\ \ \mathbf{I}]$ and $\mathbf{x} = [\mathbf{u}^H\ \ \mathbf{y}^H]^H$. The covariance matrix of $\mathbf{x}$ takes the form

$$\Lambda_{\mathbf{X}} = \langle\mathbf{x}\mathbf{x}^H\rangle = \begin{bmatrix} \Lambda_{\mathbf{U}} & \Lambda_{\mathbf{U}}\mathbf{D}^H \\ \mathbf{D}\Lambda_{\mathbf{U}} & \mathbf{D}\Lambda_{\mathbf{U}}\mathbf{D}^H + \Lambda_{\mathbf{V}} \end{bmatrix}, \tag{29.66}$$

and its inverse is equal to [40]

$$\Lambda_{\mathbf{X}}^{-1} = \begin{bmatrix} \Lambda_{\mathbf{U}}^{-1} + \mathbf{D}^H\Lambda_{\mathbf{V}}^{-1}\mathbf{D} & -\mathbf{D}^H\Lambda_{\mathbf{V}}^{-1} \\ -\Lambda_{\mathbf{V}}^{-1}\mathbf{D} & \Lambda_{\mathbf{V}}^{-1} \end{bmatrix}. \tag{29.67}$$

Substituting Eqs. (29.66) and (29.67) into Eqs. (29.62), (29.63), and (29.64), we obtain

$$\begin{aligned} F(\theta/\theta^k) &= \log|\Lambda_{\mathbf{U}}| + \log|\Lambda_{\mathbf{V}}| + tr\left\{\left(\Lambda_{\mathbf{U}}^{-1} + \mathbf{D}^H\Lambda_{\mathbf{V}}^{-1}\mathbf{D}\right)\Lambda_{\mathbf{U}|\mathbf{y}}^k\right\} \\ &+ \mu_{\mathbf{U}|\mathbf{y}}^{(k)\mathbf{H}}\left(\Lambda_{\mathbf{U}}^{-1} + \mathbf{D}^H\Lambda_{\mathbf{V}}^{-1}\mathbf{D}\right)\mu_{\mathbf{U}|\mathbf{y}}^k \\ &- 2\mathbf{y}^H\Lambda_{\mathbf{V}}^{-1}\mathbf{D}\mu_{\mathbf{U}|\mathbf{y}}^k + \mathbf{y}^H\Lambda_{\mathbf{V}}^{-1}\mathbf{y}\,, \end{aligned} \tag{29.68}$$

where

$$\mu_{\mathbf{U}|\mathbf{y}}^k = \Lambda_{\mathbf{U}}^k\mathbf{D}^{(k)H}\left(\mathbf{D}^k\Lambda_{\mathbf{U}}^k\mathbf{D}^{(k)H} + \Lambda_{\mathbf{V}}^k\right)^{-1}\mathbf{y}\,, \tag{29.69}$$

and

$$\Lambda_{\mathbf{U}|\mathbf{y}}^k = \Lambda_{\mathbf{U}}^k - \Lambda_{\mathbf{U}}^k\mathbf{D}^{(k)H}\left(\mathbf{D}^k\Lambda_{\mathbf{U}}^k\mathbf{D}^{(k)H} + \Lambda_{\mathbf{V}}^k\right)^{-1}\mathbf{D}^k\Lambda_{\mathbf{U}}^k\,. \tag{29.70}$$

Due to the constraints on the unknown parameters described in the subsection Eq. (29.62) can be written in the discrete frequency domain, as follows

$$\begin{aligned} F(\theta/\theta^k) &= N^2\log\sigma_{\mathbf{V}}^2 \\ &+ \frac{1}{\sigma_{\mathbf{V}}^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\left\{|\Delta(m,n)|^2\left(S_{\mathbf{U}|\mathbf{y}}^k(m,n) + \frac{1}{N^2}|M_{\mathbf{U}|\mathbf{y}}^k(m,n)|^2\right)\right. \\ &+ \left.\frac{1}{N^2}\left(|Y(m,n)|^2 - 2Re\left[Y^*(m,n)\Delta(m,n)M_{\mathbf{U}|\mathbf{y}}^k(m,n)\right]\right)\right\} \\ &+ \sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\left\{\log S_{\mathbf{U}}(m,n) + \frac{1}{S_{\mathbf{U}}(m,n)}\left(S_{\mathbf{U}|\mathbf{y}}^k(m,n)\right.\right. \\ &+ \left.\left.\frac{1}{N^2}|M_{\mathbf{U}|\mathbf{y}}^k(m,n)|^2\right)\right\} \end{aligned} \tag{29.71}$$

where

$$M^k_{\mathbf{U}|\mathbf{y}}(m,n) = \frac{\Delta^{(k)*}(m,n)S^k_{\mathbf{U}}(m,n)}{|\Delta^k(m,n)|^2 S^k_{\mathbf{U}}(m,n) + \sigma^{2(p)}_{\mathbf{V}}} Y(m,n)\,, \tag{29.72}$$

$$S^k_{\mathbf{U}|\mathbf{y}}(m,n) = \frac{S^k_{\mathbf{U}}(m,n)\sigma^{2(k)}_{\mathbf{V}}}{|\Delta^k(m,n)|^2 S^k_{\mathbf{U}}(m,n) + \sigma^{2(k)}_{\mathbf{V}}}\,. \tag{29.73}$$

In Eq. (29.71), $Y(m,n)$ is the 2D DFT of the observed image $y(i,j)$ and $M^k_{\mathbf{U}|\mathbf{y}}(m,n)$ is the 2D DFT of the unstacked vector $\mu^k_{\mathbf{U}|\mathbf{y}}$ into an $N \times N$ array. Taking the partial derivatives of $F(\theta/\theta^k)$ with respect to $S_{\mathbf{U}}(m,n)$ and $\Delta(m,n)$ and setting them equal to zero, we obtain the solutions that minimize $F(\theta/\theta^k)$, which represent $S^{(k+1)}_{\mathbf{U}}(m,n)$ and $\Delta^{(k+1)}(m,n)$. They are equal to

$$S^{(k+1)}_{\mathbf{U}}(m,n) = S^k_{\mathbf{U}|\mathbf{y}}(m,n) + \frac{1}{N^2}|M^k_{\mathbf{U}|\mathbf{y}}(m,n)|^2\,, \tag{29.74}$$

$$\Delta^{(k+1)}(m,n) = \frac{1}{N^2}\frac{Y(m,n)M^{(k)*}_{\mathbf{U}|\mathbf{y}}(m,n)}{S^k_{\mathbf{U}|\mathbf{y}}(m,n) + \frac{1}{N^2}|M^k_{\mathbf{U}|\mathbf{y}}(m,n)|^2}, \tag{29.75}$$

where $M^k_{\mathbf{U}|\mathbf{y}}(m,n)$ and $S^k_{\mathbf{U}|\mathbf{y}}(m,n)$ are computed by Eqs. (29.72) and (29.73). Substituting Eq. (29.75) into Eq. (29.71) and then minimizing $F(\theta/\theta^k)$ with respect to $\sigma^2_{\mathbf{V}}$, we obtain

$$\begin{aligned}\sigma^{2(k+1)}_{\mathbf{V}} &= \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\left\{|\Delta^{(k+1)}(m,n)|^2\left(S^k_{\mathbf{U}|\mathbf{y}}(m,n) + \frac{1}{N^2}|M^k_{\mathbf{U}|\mathbf{y}}(m,n)|^2\right)\right.\\ &+ \left.\frac{1}{N^2}\left(|Y(m,n)|^2 - 2Re\left[Y^*(m,n)\Delta^{(k+1)}(m,n)M^k_{\mathbf{U}|\mathbf{y}}(m,n)\right]\right)\right\}\,. \end{aligned} \tag{29.76}$$

According to Eq. (29.72) the restored image (i.e., $M^k_{\mathbf{U}|\mathbf{y}}(m,n)$) is the output of a Wiener filter, based on the available estimate of θ, with the observed image as input.

***{u,v} as the complete data* (CD_uv *algorithm*)**

The second choice of the complete data is $\mathbf{x} = [\mathbf{u}^H \;\; \mathbf{v}^H]^H$, therefore, $\mathbf{H} = [\mathbf{D}\;\; \mathbf{I}]$. Following similar steps as in the previous case it has been shown that the equations for evaluating the spectrum of the original image are the same as in the previous case, i.e., Eqs. (29.72), (29.73) and (29.74) hold true. The other two unknowns, i.e., the variance of the additive noise and the DFT of the PSF are given by

$$\sigma^{2(k+1)}_{\mathbf{V}} = \frac{1}{N^2}\sum_{m=0}^{N-1}\sum_{n=0}^{N-1}\left(S^k_{\mathbf{V}|\mathbf{y}}(m,n) + \frac{1}{N}|M^k_{\mathbf{V}|\mathbf{y}}(m,n)|^2\right)\,, \tag{29.77}$$

where

$$M^k_{\mathbf{V}|\mathbf{y}}(m,n) = \frac{\sigma^{2(k)}_{\mathbf{V}}}{|\Delta^k(m,n)|^2 S^k_{\mathbf{U}}(m,n) + \sigma^{2(k)}_{\mathbf{V}}} Y(m,n)\,, \tag{29.78}$$

$$S^k_{\mathbf{V}|\mathbf{y}}(m,n) = \frac{|\Delta^k(m,n)|^2 S^k_{\mathbf{U}}(m,n)\sigma^{2(k)}_{\mathbf{V}}}{|\Delta^k(m,n)|^2 S^k_{\mathbf{U}}(m,n) + \sigma^{2(k)}_{\mathbf{V}}}\,, \tag{29.79}$$

and

$$|\Delta^k(m,n)|^2 = \begin{cases} \frac{\frac{1}{N^2}|Y(m,n)|^2 - \sigma^{2(k)}_{\mathbf{V}}}{S^k_{\mathbf{U}}(m,n)}\,, & \text{if } \frac{1}{N^2}|Y(m,n)|^2 > \sigma^{2(k)}_{\mathbf{V}} \\ 0, & \text{otherwise}\,. \end{cases} \tag{29.80}$$

From Eq. (29.80) we observe that only the magnitude of $\Delta^k(m,n)$ is available, as was mentioned earlier. A similar observation can be made for Eq. (29.75), according to which the phase of $\Delta(m,n)$ is equal to the phase of $\Delta^0(m,n)$.

In deriving the above expressions the set of unknown parameters θ was divided into two sets $\theta_1 = \{\Lambda_{\mathbf{U}}, \Lambda_{\mathbf{V}}\}$ and $\theta_2 = \{\mathbf{D}\}$. $F(\theta_1/\theta^k)$ was then minimized with respect to θ_1, resulting in Eqs. (29.74) and (29.77). The likelihood function in Eq. (29.59) was then minimized directly with respect to $\Delta(m,n)$ assuming knowledge of θ_1^k, resulting in Eq. (29.80). The effect of mixing the optimization procedure into the EM algorithm has not been completely analyzed theoretically. That is, the convergence properties of the EM algorithm do not necessarily hold, although the application of the resulting equations increases the likelihood function. Based on the experimental results, the algorithm derived in this section always converges to a stationary point. Furthermore, the results are comparable to the ones obtained with the **CD_uy** algorithm.

{ Dx,v } as the complete data (CD_Dx,v algorithm)

The third choice of the complete data is $\mathbf{x} = [(\mathbf{Du})^H, \mathbf{v}^H]^H$. In this case, $\mathbf{D}$ and $\mathbf{x}$ cannot be estimated separately, since various combinations of $\mathbf{D}$ and $\mathbf{u}$ can result in the same $\mathbf{Du}$. The two quantities $\mathbf{D}$ and $\mathbf{u}$ are lumped into one quantity $\mathbf{t} = \mathbf{Du}$.

Following similar steps as in the two previous cases it has been shown [38, 39, 57] that the variance of the additive noise is computed according to Eq. (29.77), while the spectrum of the noise-free but blurred image $\mathbf{t}$ by the iterations

$$S_{\mathbf{T}}^{(k+1)}(m,n) = S_{\mathbf{T|y}}^{k}(m,n) + \frac{1}{N^2}|M_{\mathbf{T|y}}^{k}(m,n)|^2, \tag{29.81}$$

where

$$M_{\mathbf{T|y}}^{k}(m,n) = \frac{S_{\mathbf{T}}^{k}(m,n)}{S_{\mathbf{T}}^{k}(m,n) + \sigma_{\mathbf{V}}^{2(k)}} Y(m,n), \tag{29.82}$$

and

$$S_{\mathbf{T|y}}^{k}(m,n) = S_{\mathbf{T}}^{k}(m,n) - \frac{S_{\mathbf{T}}^{(k)2}(m,n)}{S_{\mathbf{T}}^{k}(m,n) + \sigma_{\mathbf{V}}^{2(k)}} Y(m,n). \tag{29.83}$$

Iterative Wiener Filtering

In this subsection, we deviate somewhat from the original formulation of the identification problem by assuming that the blur function is known. The problem at hand then is the restoration of the noisy-blurred image. Although there are a great number of approaches that can be followed in this case, the Wiener filtering approach represents a commonly used choice. However, in Wiener filtering knowledge of the power spectrum of the original image ($\mathbf{S_U}$) and the additive noise ($\mathbf{S_V}$) is required. A standard assumption is that of ergodicity, i.e., ensemble averages are equal to spatial averages. Even in this case, the estimation of the power spectrum of the original image has to be based on the observed noisy-blurred image, since the original image is not available. Assuming that the noise is white, its variance $\sigma_{\mathbf{V}}^2$ needs also to be estimated from the observed image. Approaches, according to which the power spectrum of the original image is computed from images with similar statistical properties, have been suggested in the literature [35]. However, a reasonable idea is to successively use the Wiener-restored image as an improved prototype for updating the unknown S_U and $\sigma_{\mathbf{V}}^2$. This idea is precisely implemented by the **CD_uy** algorithm.

More specifically, now that the blur function is known, Eq. (29.75) is removed from the EM iterations. Thus, Eqs. (29.74) and (29.76) are used to estimate $\mathbf{S_U}$ and $\sigma_{\mathbf{V}}^2$, respectively, while Eq. (29.72) is used to compute the Wiener-filtered image. The starting point $\mathbf{S_U}^0$ for the Wiener iteration can be chosen to be equal to

$$S_{\mathbf{U}}^{0}(m,n) = \hat{S}_{\mathbf{Y}}(m,n)\ , \tag{29.84}$$

where $\hat{S}_{\mathbf{Y}}(m, n)$ is an estimate of the power spectral density of the observed image. The value of $\sigma_{\mathbf{V}}^{2(0)}$ can be determined from flat regions in the observed image, since this represents a commonly used approach for estimating the noise variance.

29.4.2 Multi-Channel Image Identification and Restoration

Introduction

We use the term *multi-channel images* to define the multiple image planes (channels) which are typically obtained by an imaging system that measures the same scene using multiple sensors. Multi-channel images exhibit strong between-channel correlations. Representative examples are multispectral images [41], microwave radiometric images [42], and image sequences [43]. In the first case such images are acquired for remote sensing and facilities/military surveillance applications. The channels are the different frequency bands (color images represent a special case of great interest). In the last case the channels are the different time frames after motion compensation. More recent applications of multi-channel filtering theory include the processing of the wavelet decomposed single-channel image [44] and the reconstruction of a high resolution image from multiple low resolution images [45, 46, 47, 48].

Although the problem of single channel image restoration has been thoroughly researched, significantly less work has been done on the problem of multi-channel restoration. The multi-channel formulation of the restoration problem is necessary when cross-channel degradations exist. It can be useful, however, in the case when only within-channel degradations exist, since cross-correlation terms are exploited to achieve better restoration results [49, 50]. The cross-channel degradations may come in the form of channel crosstalks, leakage in detectors, and spectral blurs [51]. Work on restoring multi-channel images is reported in [42, 49, 50, 51, 52, 53, 54, 55], when the within- and cross-channel (where applicable) blurs are known.

29.4.3 Problem Formulation

The degradation process is modeled again as [35]

$$\mathbf{y} = \mathbf{D}\mathbf{u} + \mathbf{v} , \tag{29.85}$$

where $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{v}$ are the observed (noisy and degraded) image, the original undistorted image, and the noise process, respectively, all of which have been lexicographically ordered, and $\mathbf{D}$ the resulting degradation matrix. The noise process is assumed to be white Gaussian, independent of $\mathbf{u}$.

Let P be the number of channels, each of size $N \times N$. If $\mathbf{u}_i$, $i = 0, 1, \ldots, P-1$, represents the i-th channel. Then using the ordering of [56], the multichannel image $\mathbf{u}$ can be represented in vector form as

$$\mathbf{u} = \left[u_1(0) u_2(0) \ldots u_P(0) u_1(1) \ldots u_P(1) \ldots u_1(N^2-1) \ldots u_P(N^2-1) \right]^T . \tag{29.86}$$

Defining $\mathbf{y}$ and $\mathbf{v}$ similarly to that of Eq. (29.86), we can now use the degradation model of Eq. (29.85), recognizing that $\mathbf{y}$, $\mathbf{u}$, and $\mathbf{v}$ are of size $PN^2 \times 1$, and $\mathbf{D}$ is of size $PN^2 \times PN^2$.

Assuming that the distortion system is linear shift invariant, $\mathbf{D}$ is a $PN^2 \times PN^2$ matrix of the form

$$\mathbf{D} = \begin{bmatrix} \mathbf{D}(0) & \mathbf{D}(1) & \cdots & \mathbf{D}(N^2-1) \\ \mathbf{D}(N^2-1) & \mathbf{D}(0) & \cdots & \mathbf{D}(N^2-2) \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{D}(1) & \mathbf{D}(2) & \cdots & \mathbf{D}(0) \end{bmatrix} , \tag{29.87}$$

where the $P \times P$ sub-matrices (sub-blocks) have the form

$$\mathbf{D}(m) = \begin{bmatrix} D_{11}(m) & D_{12}(m) & \cdot\cdot & D_{1P}(m) \\ D_{21}(m) & D_{22}(m) & \cdot\cdot & D_{2P}(m) \\ \vdots & \vdots & \cdot\cdot & \vdots \\ D_{P1}(m) & D_{P2}(m) & \cdot\cdot & D_{PP}(m) \end{bmatrix}, \quad 0 \leq m \leq N^2 - 1 . \tag{29.88}$$

Note that $D_{ii}(m)$ represents the intrachannel blur, while $D_{ij}(m), i \neq j$ represents the interchannel blur. The matrix $\mathbf{D}$ in Eq. (29.87) is circulant at the block level. However, for $\mathbf{D}$ to be block-circulant, each of its subblocks $\mathbf{D}(m)$ also needs to be circulant, which, in general, is not the case. Matrices of this form are called semiblock circulant (SBC) matrices [56]. The singular values of such matrices can be found with the use of the discrete Fourier transform (DFT) kernels. Equation (29.85) can therefore be written in the vector DFT domain [56].

Similarly, the covariance matrix of the original signal, $\Lambda_{\mathbf{U}}$, and the covariance matrix of the noise process, $\Lambda_{\mathbf{V}}$, are also semiblock circulant (assuming $\mathbf{u}$ and $\mathbf{v}$ are stationary). Note that $\Lambda_{\mathbf{U}}$ is not block-circulant because there is no justification to assume stationarity between channels (i.e., $\Lambda_{\mathbf{U}_i\mathbf{U}_j}(m) = E[\mathbf{u}_i(m)\mathbf{u}_j(m)^*]$ is not equal to $\Lambda_{\mathbf{U}_{i+p}\mathbf{U}_{j+p}}(m) = E[\mathbf{u}_{i+p}(m)\mathbf{u}_{j+p}(m)^*]$ [50], where $\Lambda_{\mathbf{U}_i\mathbf{U}_j}(m)$ is the $(i,j)^{th}$ submatrix of $\Lambda_{\mathbf{U}}$). However, $\Lambda_{\mathbf{U}}$ and $\Lambda_{\mathbf{V}}$ are semiblock circulant because $\mathbf{u}_i$ and $\mathbf{v}_i$ are assumed to be stationary within each channel.

29.4.4 The E-Step

We follow here similar steps to the ones presented in the previous section. We choose $[\mathbf{u}^H\mathbf{y}^H]^H$ as the complete data. Since the matrices Λ_U, Λ_V, and D, are assumed to be semi-block circulant, the E-step requires the evaluation of

$$F\left(\theta;\theta^k\right) = \sum_{m=0}^{N-1}\sum_{n=0}^{N-1} J(m,n) , \tag{29.89}$$

where

$$\begin{aligned} J(m,n) &= \log|\Theta_{\mathbf{U}}(m,n)| + \log|\Theta_{\mathbf{V}}(m,n)| + \mathrm{tr}\Big\{\Big(\Theta_{\mathbf{U}}^{-1}(m,n) \\ &+ \Theta_{\mathbf{D}}^H(m,n)\Theta_{\mathbf{V}}^{-1}(m,n)\Theta_{\mathbf{D}}(m,n)\Big)\,\Theta_{\mathbf{U}|\mathbf{y}}^k(m,n)\Big\} \\ &+ \frac{1}{N^2}\mathrm{tr}\Big\{\Big[\Theta_{\mathbf{U}}^{-1}(m,n) + \Theta_{\mathbf{D}}^H(m,n)\Theta_{\mathbf{V}}^{-1}(m,n)\Theta_{\mathbf{D}}(m,n)\Big] \\ &\times \mathbf{M}_{\mathbf{U}|\mathbf{y}}^k(m,n)\mathbf{M}_{\mathbf{U}|\mathbf{y}}^{(k)H}(m,n)\Big\} \\ &- \frac{1}{N^2}\Big(\mathbf{Y}^H(m,n)\Theta_{\mathbf{V}}^{-1}(m,n)\Theta_{\mathbf{D}}(m,n)\,\mathbf{M}_{\mathbf{U}|\mathbf{y}}^k(m,n) \\ &+ \mathbf{M}_{\mathbf{U}|\mathbf{y}}^{(k)H}(m,n)\Theta_{\mathbf{D}}^H(m,n)\Theta_{\mathbf{V}}^{-1}(m,n)\mathbf{Y}(m,n)\Big) \\ &+ \frac{1}{N^2}\mathbf{Y}^H(m,n)\Theta_{\mathbf{V}}^{-1}(m,n)\mathbf{Y}(m,n) . \end{aligned} \tag{29.90}$$

The derivation of Eq. (29.90) is presented in detail in [48, 57, 58]. Equation (29.89) is the corresponding equation to Eq. (29.71) for the multichannel case.

In Eq. (29.90), $\Theta_{\mathbf{U}}(m,n)$ is the (m,n)-th component matrix of $\Theta_{\mathbf{U}}$, which is related to $\Lambda_{\mathbf{U}}$ by a similarity transformation using two-dimensional discrete Fourier kernels [56, 57]. To be more specific, for $P = 3$, the matrix,

$$\Theta_{\mathbf{U}}(m,n) = \begin{bmatrix} S_{11}(m,n) & S_{12}(m,n) & S_{13}(m,n) \\ S_{21}(m,n) & S_{22}(m,n) & S_{23}(m,n) \\ S_{31}(m,n) & S_{32}(m,n) & S_{33}(m,n) \end{bmatrix}, \tag{29.91}$$

consists of all the (m,n)-th component of the power and cross power spectra of the original color image (without loss of generality in the subsequent discussion three-channel examples will be used). It is worthwhile noting here that the power spectra $S_{ii}(m,n)$, $i = 1, 2, 3$, which are the diagonal entries of $\Theta_{\mathbf{U}}(m,n)$, are real-valued, while the cross power spectra (the off-diagonal entries) are complex. This illustrates one of the main differences between working with multichannel images as opposed to single-channel images. In addition to each frequency component being a $P \times P$ matrix versus a scalar quantity for the single-channel case, the cross power spectra is complex versus being real for the single-channel case. Similarly, the (m,n)-th component of the inverse of the noise spectrum matrix is given by

$$\Theta_{\mathbf{V}}{}^{-1}(m,n) = \begin{bmatrix} z_{11}(m,n) & z_{12}(m,n) & z_{13}(m,n) \\ z_{21}(m,n) & z_{22}(m,n) & z_{23}(m,n) \\ z_{31}(m,n) & z_{32}(m,n) & z_{33}(m,n) \end{bmatrix}. \tag{29.92}$$

One simplifying assumption that we can make about Eq. (29.92) is that the noise is white within channels and zero across channels. This results in $\Theta_{\mathbf{V}}(m,n)$ being the same diagonal matrix for all (m,n).

$\Theta_{\mathbf{D}}(m,n)$ in Eq. (29.90) is equal to

$$\Theta_{\mathbf{D}}(m,n) = \begin{bmatrix} \Delta_{11}(m,n) & \Delta_{12}(m,n) & \Delta_{13}(m,n) \\ \Delta_{21}(m,n) & \Delta_{22}(m,n) & \Delta_{23}(m,n) \\ \Delta_{31}(m,n) & \Delta_{32}(m,n) & \Delta_{33}(m,n) \end{bmatrix}, \tag{29.93}$$

where $\Delta_{ij}(m,n)$ is the within-channel $(i = j)$ or cross-channel $(i \neq j)$ frequency response of the blur system, and $\mathbf{Y}(m,n)$ is the (m,n)-th component of the DFT of the observed image. $\Theta^k_{\mathbf{U}|\mathbf{y}}(m,n)$ and $\mathbf{M}^k_{\mathbf{U}|\mathbf{y}}(m,n)$ are the (m,n)-th frequency component matrix and vector of the multichannel counterparts of $\Lambda_{\mathbf{U}|\mathbf{y}}$ and $\mu_{\mathbf{U}|\mathbf{y}}$, respectively, computed by

$$\begin{aligned} \Theta^k_{\mathbf{U}|\mathbf{y}}(m,n) &= \Theta^k_{\mathbf{U}}(m,n) - \Theta^k_{\mathbf{U}}(m,n)\Theta^{(k)H}_{\mathbf{D}}(m,n)\Big[\Theta^k_{\mathbf{V}}(m,n) \\ &+ \Theta^k_{\mathbf{D}}(m,n)\Theta^k_{\mathbf{U}}(m,n)\Theta^{(k)H}_{\mathbf{D}}(m,n)\Big]^{-1}\Theta^k_{\mathbf{D}}(m,n)\Theta^k_{\mathbf{U}}(m,n) \end{aligned} \tag{29.94}$$

and

$$\begin{aligned} \mathbf{M}^k_{\mathbf{U}|\mathbf{y}}(m,n) &= \Theta^k_{\mathbf{U}}(m,n)\Theta^{(k)H}_{\mathbf{D}}(m,n)\Big[\Theta^k_{\mathbf{V}}(m,n) \\ &+ \Theta^k_{\mathbf{D}}(m,n)\Theta^k_{\mathbf{U}}(m,n)\Theta^{(k)H}_{\mathbf{D}}(m,n)\Big]^{-1}\mathbf{Y}(m,n). \end{aligned} \tag{29.95}$$

29.4.5 The M-Step

The M-step requires the minimization of $J(m,n)$ with respect to $\Theta_{\mathbf{U}}(m,n)$, $\Theta_{\mathbf{V}}(m,n)$ and $\Theta_{\mathbf{D}}(m,n)$. The resulting solutions become $\Theta^{(k+1)}_{\mathbf{U}}(m,n)$, $\Theta^{(k+1)}_{\mathbf{V}}(m,n)$ and $\Theta^{(k+1)}_{\mathbf{D}}(m,n)$, respectively.

The minimization of $J(m,n)$ with respect to $\Theta_{\mathbf{U}}$ is straightforward, since $\Theta_{\mathbf{U}}$ is decoupled from $\Theta_{\mathbf{V}}(\mathbf{m},\mathbf{n})$ and $\Theta_{\mathbf{D}}$. An equation similar to Eq. (29.74) results. The minimization of $J(m,n)$ with respect to $\Theta_{\mathbf{D}}$ is not as straightforward; $\Theta_{\mathbf{D}}$ is coupled with $\Theta_{\mathbf{V}}$. Therefore, in order to minimize $J(m,n)$ with respect to $\Theta_{\mathbf{D}}$, $\Theta_{\mathbf{V}}$ must be solved first in terms of $\Theta_{\mathbf{D}}$, substituted back into Eq. (29.90), and then minimized with respect to $\Theta_{\mathbf{D}}$.

It is shown in [48, 58] that two conditions must be met in order to obtain explicit equations for the blur. First, the noise spectrum matrix, $\Theta_{\mathbf{V}}(m,n)$, must be a diagonal matrix, which is frequently encountered in practice. Second, all of the blurs must be symmetric, so that there is no phase when working in the discrete frequency domain. The first condition arises from the fact that $\Theta_{\mathbf{V}}(m,n)$ and $\Theta_{\mathbf{D}}(m,n)$ are coupled. The second condition arises from the Cauchy-Riemann theorem, and must be satisfied in order to guarantee the existence of a derivative at every point.

With these conditions, the iterations for $\Delta(m,n)$ and $\sigma_{\mathbf{V}}(m,n)$ are derived in [48, 58], which are similar respectively to Eqs. (29.75) and (29.76). Special cases are also analyzed in [48, 58], when the number of unknowns is reduced. For example, if $\Theta_{\mathbf{D}}$ is known, the multichannel Wiener filter results.

29.5 Experimental Results

The effectiveness of both the single channel and multi-channel restoration and identification algorithms is demonstrated experimentally. The red, green, and blue (RGB) channels of the original Lena image used for this experiment are shown in Fig. 29.1. A 5×5 truncated Gaussian blur is used for each channel and Gaussian white noise is added resulting in a blurred signal-to-noise ratio (SNR) of 20 dB. The degraded channels are shown in Fig. 29.2. Three different experiments were performed with the available degraded data. The single channel algorithm of Eqs. (29.74), (29.75), and (29.76) was first run for each of the RGB channels independently. The restored images are shown in Fig. 29.3. The corresponding multichannel algorithm was then run, resulting in the restored channels shown in Fig. 29.4. Finally the multichannel Wiener filter was also run, in demonstrating the upper bound of the algorithm's performance, since the blurs are now exactly known. The resulting restored images are shown in Fig. 29.5. The improvement in SNR for the three experiments and for each channel is shown in Table 29.1. According to this table, the performance of the algorithm increases from the first

TABLE 29.1 Improvement in SNR (dB)

η	Decoupled EM	Multichannel EM	Wiener
Red	1.5573	2.1020	2.3420
Green	1.3814	2.0086	2.3181
Blue	1.1520	1.5148	1.8337

to the last experiment. This is to be expected, since in considering the multichannel algorithm over the single channel algorithm the correlation between channels is taken into account, which brings additional information into the problem.

A photographically blurred image is shown next in Fig. 29.6. The restorations of it by the **CD_uy** and **CD_uv** algorithms are shown, respectively, in Figs. 29.7 and 29.8.

29.5.1 Comments on the Choice of Initial Conditions

The likelihood function which is optimized is highly nonlinear and a number of local minima exist. Although the incorporation of the various constraints, discussed earlier, restricts the set of possible solutions, a number of local minima still exist. Therefore, the final result depends on the initial conditions. Based on our experience in implementing the EM iterations of the previous sections for the single-channel and the multi-channel image restoration cases, the following comments and

FIGURE 29.1: Original RGB Lena.

FIGURE 29.2: Degraded RGB Lena, intra-channel blurs only, 20 dB SNR.

FIGURE 29.3: Restored RGB by the decoupled single channel EM algorithm.

observations are in order.

It was observed experimentally that the final results are quite insensitive to variations in the values of the noise variance(s) and the original image power spectra. An estimate of the noise variances from flat regions of the noisy and blurred images were used as initial condition. It was observed that using initial estimates of the noise variances larger than the actual ones produced good final results.

The final results are quite sensitive, however, to variations in the values of the PSF. Knowledge of the support of the PSF is quite important. In [38] after convergence of the EM algorithm the estimate of the PSF was truncated, normalized, and used as an initial condition in restarting another iteration cycle.

29.6 Summary and Conclusion

In this chapter, we have described and illustrated how the EM algorithm can be used in image recovery problems. The basic approach can be summarized by the following steps.

FIGURE 29.4: Restored RGB Lena by the multi-channel EM algorithm.

FIGURE 29.5: Restored RGB Lena by the iterative multi-channel Wiener algorithm.

FIGURE 29.6: Photographically blurred image.

FIGURE 29.7: Restored image by the **CD_uy** algorithm.

FIGURE 29.8: Restored image by the **CD_uv** algorithm.

1. Select a statistical model for the observed data and formulate the image recovery problem as an MLE problem.
2. If the likelihood function is difficult to optimize directly, the EM algorithm can be used by properly selecting the complete data.
3. Constraints on the parameters or image to be estimated, proper initial conditions, and multiple complete data spaces can be considered to improve the uniqueness and convergence of the estimates.
4. Derive the equations for the E-step and M-step.

We end this chapter with several remarks. We want to emphasize again that the EM algorithm only guarantees convergence to a local optimum. Therefore, the initial conditions are quite critical, as is also discussed in the previous section. Depending on the number of the unknown parameters, one could consider evaluating in a systematic fashion the likelihood function directly at a number of points and use as initial condition the point which results in the largest value of the likelihood function. Improved results can be obtained potentially if the number of the unknown parameters is reduced by parameterizing the unknown functions. For example, separable and nonseparable exponential covariance models are used in [46, 47, 48], and an autoregressive model in [38, 57] to model the original image, and parameterized blur models are discussed in [38]. We want to mention also that the EM algorithm can be implemented in different domains. For example, it is implemented in both spatial and frequency domains, respectively, in sections 29.3 and 29.4. Other domains are also possible by applying proper transforms, e.g., the wavelet transform [59].

References

[1] Jain, A.K., *Fundamentals of Digital Image Processing*, Prentice Hall, Englewood Cliffs, NJ, 1989.

[2] Yang, Y., Galatsanos, N.P., and Katsaggelos, A.K., Regularized image reconstruction to remove blocking artifacts from block discrete cosine transform compressed images, *IEEE Trans. Circuits Syst. Video Technol.*, 3(6): 421–432, December, 1993.

[3] Yang, Y., Galatsanos, N.P., and Katsaggelos, A.K., Projection-based spatially-adaptive reconstruction of block transform compressed images, *IEEE Trans. Image Process.*, 4(7): 896–908, July, 1995.

[4] Parker, A.J., *Image Reconstruction in Radiology*, CRC Press, Boca Raton, FL, 1990.

[5] Russ, J.C., *The Image Processing Handbook*, CRC Press, Boca Raton, FL, 1992.

[6] Snyder, D.L. and Miller, M.I., *Random Processes in Time and Space*, 2nd ed., Springer-Verlag, 1991.
[7] Shepp, L. and Vardi, Y., Maximum-likelihood reconstruction for emission tomography, *IEEE Trans. Med. Imag.*, 1: 113–122, Oct., 1982.
[8] Katsaggelos, A.K., Ed., *Digital Image Restoration*, Springer-Verlag, 1991.
[9] Lagendijk, R.L. and Biemond, J., *Iterative Identification and Restoration of Images*, Kluwer Academic Publishers, 1991.
[10] Cox, D.R and Hinkley, D.V., *Theoretical Statistics*, Chapman and Hall, 1974.
[11] Dempster, A.P., Laird, N.M., and Rubin, D.B., Maximum likelihood from incomplete data via the EM algorithm, *J. Roy. Soc. Statist., Series B*, 39: 1–38, 1977.
[12] Hebert, T. and Leahy, R., A generalized EM algorithm for 3-D Bayesian reconstruction from Poisson data using Gibbs priors, *IEEE Trans. Med. Imaging*, 8: 194–202, June, 1989.
[13] Green, P.J., On use of the EM algorithm for penalized likelihood estimation, *J. Roy. Soc. Statist., Series B*, 52: 443–452, 1990.
[14] Zhang, J., The mean field theory in EM procedures for Markov random fields, *IEEE Trans. ASSP*, 40: 2570–2583, October, 1992.
[15] Zhang, J., The mean field theory in EM procedures for blind Markov random field image restoration, *IEEE Trans. Image Process.*, 2: 27–40, Jan., 1993.
[16] Feder, M. and Weinstein, E., Parameter estimation of superimposed signals using the EM algorithm, *IEEE Trans. ASSP*, 36: 477–489, April, 1988.
[17] Fessler, J.A. and Hero, A.O., Space alternating generalized expectation-maximization algorithm, *IEEE Trans. SP*, 42: 2664–2678, Oct., 1994.
[18] Fessler, J.A. and Hero, A.O., Complete data space and generalized EM algorithm, *Proc. ICASSP*, Vol. IV, pp. 1–4, Mineappolis, Minnesota, April 27-30, 1993.
[19] Hero, A.O and Fessler, J.A., Convergence in norm for alternating expectation-maximization (EM) type algorithms, *Statistica Sinica*, 5: 41–54, Jan., 1995.
[20] Wu, J., On the convergence properties of the EM algorithm, *The Annals of Statistics*, 11: 95–103, 1983.
[21] Redner, R.A. and Walker, H.F., Mixture densities, maximum likelihood and the EM algorithm, *SIAM Review*, 26(2): 195–239, 1984.
[22] Besag, J., Spatial interaction and the statistical analysis of lattice systems, *J. Roy. Statist. Soc., Series B*, 36: 192–226, 1974.
[23] Geman, S. and Geman, D., Stochastic relaxation, Gibbs distribution, and the Bayesian restoration of images, *IEEE Trans. PAMI*, 6: 721–741, Nov., 1984.
[24] Chellappa, R. and Jain, A., Eds., *Markov Random Fields — Theory and Applications*, Academic Press, 1993.
[25] Chandler, D., *Introduction to Modern Statistical Mechanics*, Oxford University Press, 1987.
[26] Bouman, C. and Sauer, K., Maximum likelihood scale estimation for a class of Markov random fields, *Proc. ICASSP*, pp. V537–540, Adelaide, Australia, April, 19-22, 1994.
[27] Metropolis, N., et al., Equation of state calculation by fast computing machines, *J. Chem. Phys.*, 21: 1087–1092, 1953.
[28] Konrad, J. and Dubois, E., Comparison of stochastic and deterministic solution methods in Bayesian estimation of 2D motion, *Image and Visual Computing*, 8(4): 304–317, Nov., 1990.
[29] Cover, T. and Thomas, J., *Elements of Information Theory*, John Wiley & Sons, 1992.
[30] Zhang, J., The application of the Gibbs-Bogoliubov-Feynmann inequality in the mean field theory for Markov random fields, Preprint, 1995.
[31] Wu, C.-H. and Doerschuk, P.C., Cluster expansions for the deterministic computation of Bayesian estimators based on Markov random fields, *IEEE Trans. PAMI*, 17: 275–293, March, 1995.

[32] Anderson, B.D.O. and Moore, J. B., *Optimal Filtering*, Prentice-Hall, Englewood Cliffs, NJ, 1979.

[33] Banham, M.R. and Katsaggelos, A.K., Digital restoration of images, *IEEE Signal Process. Mag.*, 14(2): 24–41, Mar., 1997.

[34] Tekalp, A.M., Kaufman, H., and Woods, J.W., *IEEE Trans. ASSP*-**34**: 963–972, 1986.

[35] Andrews, H.C. and Hunt, B.R., *Digital Image Restoration*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[36] Dudgeon, D.E. and Mersereau, R.M., *Multidimensional Digital Signal Processing*, Prentice-Hall, Englewood Cliffs, NJ, 1984.

[37] Gray, R.M., *IEEE Trans. IT*-**18**: 725–730, 1985.

[38] Katsaggelos, A.K. and Lay, K.T., Identification and restoration of images using the expectation maximization algorithm, in *Digital Image Restoration*, Katsaggelos, A.K., Ed., Springer-Verlag, 1991.

[39] Lay, K.T. and Katsaggelos, A.K., Image identification and restoration based on the expectation-maximization algorithm, *Opt. Eng.*, 29: 436–445, May, 1990.

[40] Kailath, T., *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[41] Lee, J.B., Woodyatt, A.S., and Berman, M., Enhancement of high spectral resolution remote-sensing data by a noise adjusted principle component transform, *IEEE Trans. Geosci. Remote Sens.*, 28(3): 295–304, 1990.

[42] Chin, R.T., Yeh, C.L., and Olson, W.S., Restoration of multichannel microwave radiometric images, *IEEE Trans. Patt. Anal. Mach. Intell.*, PAMI-7(4): 475–484, July, 1985.

[43] Choi, M.G., Galatsanos, N.P., and Katsaggelos, A.K., Multichannel regularized iterative restoration of image sequences, *J. Visual Commun. Image Represent.*, 7(3): 244–258, Sept., 1996.

[44] Banham, M.R., Galatsanos, N.P., Gonzalez, H., and Katsaggelos, A.K., Multichannel restoration of single channel images using a wavelet-based subband decomposition, *IEEE Trans. Image Process.*, 3(6): 821–833, Nov., 1994.

[45] Tsai, R.Y. and Huang, T.S., Multiframe image restoration and registration, in *Advances in Computer Vision and Registration*, vol. 1, *Image Reconstruction from Incomplete Observations*, Huang, T.S., Ed., pp. 317–339, JAI Press, 1984.

[46] Tom, B.C. and Katsaggelos, A.K., Reconstruction of a high resolution image from multiple degraded mis-registered low resolution images, *Proc. SPIE, Visual Communications and Image Processing*, Chicago, Vol. 2308, pt. 2, pp. 971–981, Sept., 1994.

[47] Tom, B.C., Katsaggelos, A.K., and Galatsanos, N.P., Reconstruction of a high resolution from registration and restoration of low resolution images, *IEEE Proc. Int. Conf. Image Process.*, Austin, Vol. 3, pp. 553–557, Nov., 1994.

[48] Tom, B.C., Reconstruction of a High Resolution Image from Multiple Degraded Mis-Registered Low Resolution Images, Ph.D. thesis, Northwestern University, Dept. of EECS, June, 1995.

[49] Hunt, B.R. and Kübler, O., Karhunen-Loeve multispectral image restoration, part I : theory, *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-32(3): 592–600, June, 1984.

[50] Galatsanos, N.P. and Chin, R.T., Digital restoration of multichannel images, *IEEE Trans. Acoust., Speech, Signal Process.*, ASSP-37(3): 415–421, March, 1989.

[51] Galatsanos, N.P. and Chin, R.T., Restoration of color images by multichannel Kalman filtering, *IEEE Trans. Signal Process.*, 39(10): 2237–2252, Oct., 1991.

[52] Galatsanos, N.P., Katsaggelos, A.K., Chin, R.T., and Hillery, A.D., Least squares restoration of multichannel images, *IEEE Trans. Signal Process.*, 39: 2222–2236, Oct., 1991.

[53] Tekalp, A.M. and Pavlovic, G., Multichannel image modeling and Kalman filtering for multi-spectral image restoration, *IEEE Trans. Signal Process.*, 19(3): 221–232, March, 1990.

[54] Kang, M.G. and Katsaggelos, A.K., Simultaneous multichannel image restoration and estimation of the regularization parameters, *IEEE Trans. Image Process.*, 6(5) 774–778, May, 1997.

[55] Zhu, W., Galatsanos, N.P., and Katsaggelos, A.K., Regularized multichannel restoration using cross-validation, *Graph. Models Image Process.*, 57(1): 38–54, Jan., 1995.
[56] Katsaggelos, A.K., Lay, K.T., and Galatsanos, N.P., A general framework for frequency domain multichannel signal processing, *IEEE Trans. Image Process.*, 2(3): 417–420, July, 1993.
[57] Lay, K.T., Blur Identification and Image Restoration Using the EM Algorithm, Ph.D. thesis, Northwestern University, Dept. of EECS, Dec., 1991.
[58] Tom, B.C.S., Lay, K.T., and Katsaggelos, A.K., Multi-channel image identification and restoration using the expectation-maximization algorithm, *Optical Engineering*, Special Issue on Visual Communications and Image Processing, 35(1): 241–254, Jan., 1996.
[59] Banham, M.R., Wavelet Based Image Restoration Techniques, Ph.D. thesis, Northwestern University, Dept. of EECS, June, 1994.